Computational Bayesian Coursework: Peak Traffic Accident Rates

Muneeb (Student #: 151244472)

1. Abstract

The goal of this report is to compare the number of accidents in peak commute working hours vs non-peak hours and see if there is a statistically significant difference between the two. This is done by using Bayesian Inference and running a STAN model on two samples with the same conjugate prior. By demonstrating two different parameter values, the method establishes that there was enough proof in the data to skew the posterior in different directions.

2. Data

The US Accidents (2016-21) is a traffic dataset available on Kaggle. First compiled in 2019, it has been updated by Moosavi et al. [1] from Ohio State University since then each year. Data was primarily pulled from MapQuest (~70%) and MS Bing Traffic Collector (~30%) APIs. These APIs broadcast events US state and federal transportation agencies, law-enforcement agencies, traffic cameras, and traffic sensors. Conservative settings described in the paper were used to ensure removal of duplicates across sources.

Once the accident data was gathered in 2019, it was updated in 2020 and 2021 separately. These updates have made the data quality in recent years better; however it also means previous year trends might not align anymore.

After integrating the datasets, the new combined dataset is augmented with other potential features ranging from geo-location to weather data. These can be important features when trying to explain what causes accident rates to be so high; however, since we are majorly concerned with if accidents counts are higher during peak commmute hours we can ignore most of the features. Therefore, in this report we use just the timestamp of when the accident happened.

2.1 Loading Data

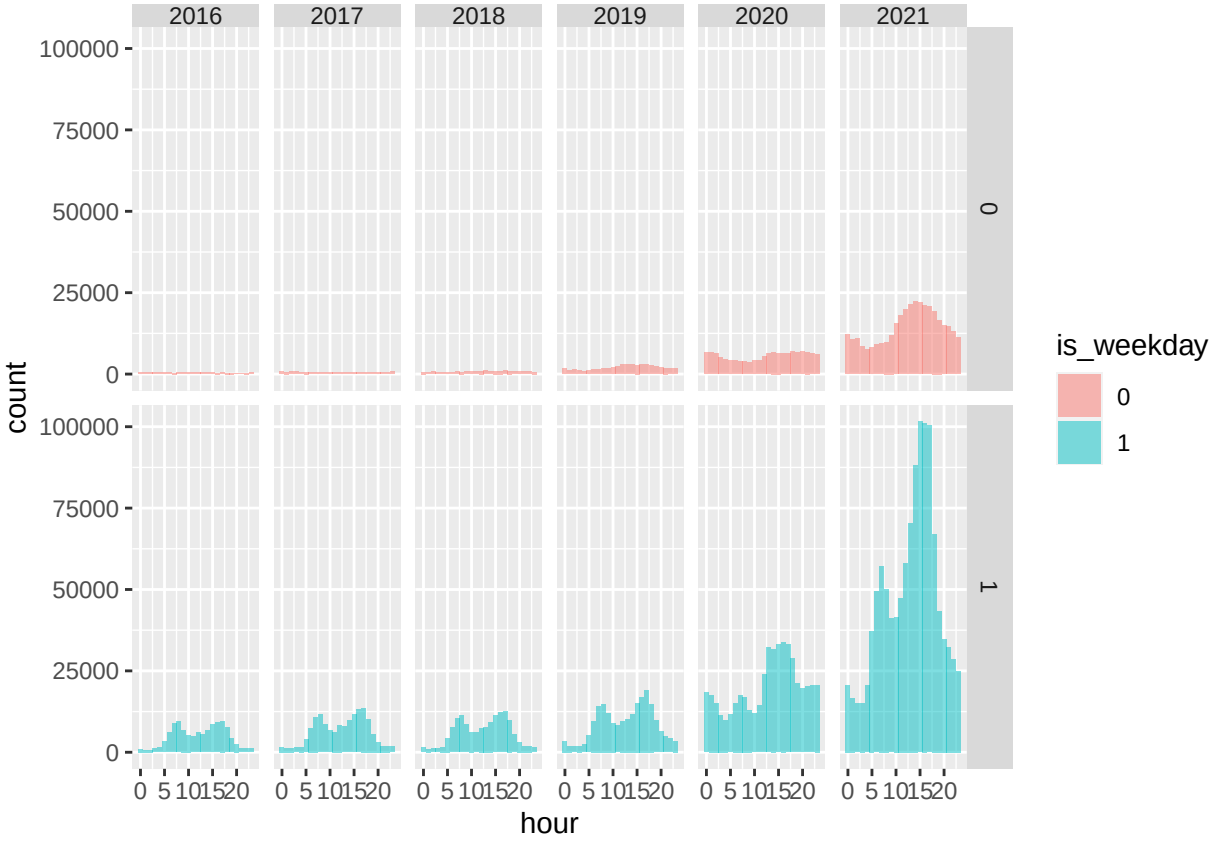
As this is a large dataset, the data can be efficiently loaded using the fread function in data.table

```
library(data.table)
setwd("/mnt/d/Grad/Period 2/Computational Bayesian")

# # You can use this piece of code to load the dataset from the kaggle csv file
# # https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download
# acc.data <- fread("US_Accidents_Dec21_updated.csv")
# colnames(acc.data) <- sapply(colnames(acc.data), tolower)
# acc.data <- acc.data[,acc_time:=ymd_hms(start_time)
#          ][,is_weekday:=fcase(wday(acc_time)%%7>1,1,default=0)
#          ][,hour:=hour(acc_time)
#          ][,day:=date(acc_time)
#          ][,year:=year(acc_time)
#          ][,.(id, severity, acc_time, is_weekday, hour, day, year)]

acc.data <- fread("sample_accidents.csv")[,.(id, acc_time, is_weekday, hour, day, year)]
acc.data |> head()
```

```
##     id            acc_time is_weekday hour        day year
```

```
## 1: A-1 2016-02-08 00:37:08          1    0 2016-02-08 2016
## 2: A-2 2016-02-08 05:56:20          1    5 2016-02-08 2016
## 3: A-3 2016-02-08 06:15:39          1    6 2016-02-08 2016
## 4: A-4 2016-02-08 06:51:45          1    6 2016-02-08 2016
## 5: A-5 2016-02-08 07:53:43          1    7 2016-02-08 2016
## 6: A-6 2016-02-08 08:16:57          1    8 2016-02-08 2016
```

2.2 Exploring Data



Observing the data, there are a few points that immediately pop out. Firstly, data before 2020 seems to be severely under-counted. Natural trends don't indicate any reason that yearly accidents should increase almost four-fold. Therefore, it is reasonable to believe that the underlying data collection was improved in 2020 and once again in 2021. This also alligns with how this data was collected, initially in 2019 up till 2019 and then updated once per year. Therefore, moving on with this data we will only be using data from 2021 (2020 can be ignored due to COVID-19 as well). Secondly, weekday and weekend trends seem to differ. Hence, we'll also only be using weekday data to try and control for outside factors.

3. Statistical Method

Looking at the data we can immediately see that the number of accidents in an hour can be modeled as a Poisson distribution with a certain mean, λ. Expanding on this idea, we can compare two distributions, one for peak commute hours and one for without. The overlap in the posterior distribution of the mean rates would indicate how likely or unlikely it is that there is a difference.

We define peak commute hours as 8-10 am and 4-6 pm. Non peak hours are between 10 am and 11 pm not included in peak hours. Midnight to 7 am is not included to avoid decreasing the accident counts in the non-peak hour sample due to general inactivity.

A STAN model will be fit on these two samples with the same prior to investigate if the underlying distribution is the same or not. STAN uses Metropolis-Hastings Algorithm to sample from the posterior

distribution. MHC works by drawing samples from a proposed distribution conditional on the previous sample. It accepts these samples as distributions from the target distribution with a ratio of the relative densities of the two distributions. In this case, since the prior was set to be a Gamma distribution, a conjugate of the Poisson distribution, there also exists a closed-form solution of this set up.

3.1 STAN Model

We sample for two parameters λ_{high} and λ_{low} of a Poisson distribution, both with the same non-informative Gamma prior.

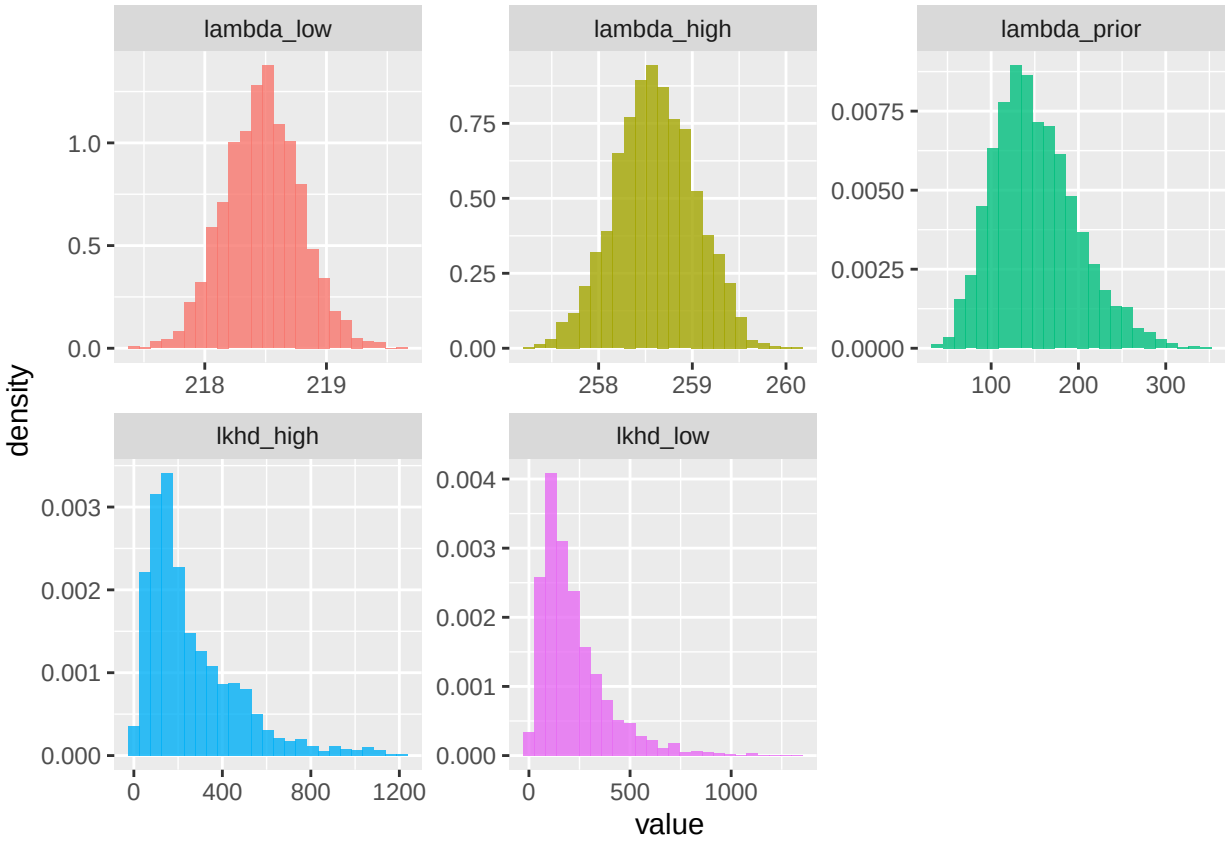
```
data {
  int<lower=1> n[2];
  int Y_low[n[1]];
  int Y_high[n[2]];
}
parameters {
  real<lower=0> lambda_high;
  real<lower=0> lambda_low;
}
model {
    Y_high ~ poisson(lambda_high);
    Y_low ~ poisson(lambda_low);
    lambda_high ~ gamma(15,15);
    lambda_low ~ gamma(15,15);
}
```

3.2 Model Sampling and Evaluation

```
library(rstan)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)

acc.data_filterd <- acc.data[year==2021 & is_weekday==1,.(.N),by=.(day, hour)]
model.data <- list(n=c(acc.data_filterd[hour %in% c(11,12,13,14,15,19,20,21,22),.N],
                       acc.data_filterd[hour %in% c(8,9,10,16,17,18),.N]),
                   Y_low=acc.data_filterd[hour %in% c(11,12,13,14,15,19,20,21,22),N],
                   Y_high=acc.data_filterd[hour %in% c(8,9,10,16,17,18),N])

acc.model <- sampling(acc.stan,
                          data=model.data,
                          chains = 4,
                          warmup = 500,
                          iter = 1000,
                          cores = 4
                        )
```

```
## Inference for Stan model: d92b3e74da6323d28353fa7924a73741.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##                   mean se_mean   sd       2.5%        25%        50%        75%
## lambda_high     258.61    0.01 0.43     257.75     258.32     258.60     258.90
## lambda_low      218.48    0.01 0.31     217.88     218.27     218.48     218.68
## lp__        4043364.75    0.04 1.08 4043361.71 4043364.34 4043365.09 4043365.51
```

```
##                   97.5% n_eff Rhat
## lambda_high      259.44  1510    1
## lambda_low       219.10  2020    1
## lp__         4043365.79   745    1
##
## Samples were drawn using NUTS(diag_e) at Tue Dec 13 08:07:32 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```



The model shows a significant difference between λ_{low} and λ_{high}. The 95% credible interval for λ_{low} is between 216.04 and 217.13 whereas the 95% CI for λ_{high} is 257.82 to 259.39. With such a huge difference w.r.t. the deviations in each parameter, we can confidently say that accidents during peak commute hours happen with a higher mean rate than accidents in non-peak commute hours. In fact if we use the posterior samples that STAN generates and do a Monte Carlo simulation all 100% of λ_{high} are greater than λ_{low}.

Furthermore, the Rhat value in STAN indicates that each chain converged i.e. for four chains, four random starting points were picked and by the end of the iteration, there was no way to know which value came from which chain. This further increases the confidence in the result presented in this report.

References

- [1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.